Retrieving relatives from historical data

Hundt, Marianne ; Denison, David ; Schneider, Gerold

Abstract: Variation and change in relativization strategies has been well documented (e.g. Ball 1996: 46, Biber and Clark 2002, Biber, Johansson, Leech, Conrad and Finegan 1999, Johansson 2006, Lehmann 2002). Certain types of relative clause, namely that-relatives and zero relatives, were difficult to retrieve from plain-text corpora. Studies therefore either relied on manual extraction of data or a subset of possible relativization strategies. In some text types, however, the zero relative is an important member of the class of possible relativizers. Recent advances in syntactic annotation should have made that-relatives and zero relatives more accessible to automatic retrieval. In this article, we test precision and recall of searches on a modest-sized corpus, i.e. scientific texts from ARCHER (A Representative Corpus of Historical English Registers), as a preliminary to future work on the large corpora which are increasingly becoming available. The parser retrieved some false positives and at the same time missed some relevant data. We discuss structural reasons for both kinds of shortcoming as well as the possibilities and limitations of parser adaptation.

DOI: https://doi.org/10.1093/llc/fqr049

Retrieving relatives from historical data

Marianne Hundt, Zürich, David Denison, Manchester and Gerold Schneider, Zürich

corresponding author:
Prof. Dr. Marianne Hundt
Englisches Seminar
Plattenstrasse 47
CH-8032 Zürich
m.hundt@es.uzh.ch

12 Abstract

13 Variation and change in relativization strategies has been well documented (e.g. Ball 1996: 46, Biber
14 and Clark 2002, Biber, Johansson, Leech, Conrad and Finegan 1999, Johansson 2006, Lehmann 2002,
15 Sigley 1997). Certain types of relative clause, namely *that*-relatives and zero relatives, were difficult
16 to retrieve from plain-text corpora (Biber 1988, Olofsson 1981). Studies therefore either relied on
17 manual extraction of data or a subset of possible relativization strategies. In some text types, however,
18 the zero relative is an important member of the class of possible relativizers (Ball 1994). Recent
19 advances in syntactic annotation should have made *that*-relatives and zero relatives more accessible to
20 automatic retrieval. In this paper, we test precision and recall of searches on a modest-sized corpus,
21 i.e. scientific texts from ARCHER (*A Representative Corpus of Historical English Registers*), as a
22 preliminary to future work on the large corpora which are increasingly becoming available. The
23 parser retrieved some false positives and at the same time missed some relevant data. We discuss
24 structural reasons for both kinds of shortcoming as well as the possibilities and limitations of parser
25 adaptation.

26

27 Acknowledgement

28 We would like to thank the two anonymous reviewers of *Literary and Linguistic Computing* for their
29 valuable comments.

30

1. Introduction

Olofsson (1981) is an early corpus-based study that addresses the accessibility of relative clauses to automatic retrieval. His study is based on one of the first standard reference corpora of Present-Day (American) English, the Brown corpus. The corpus was available in a version that had been tagged for parts of speech (POS) but not parsed. POS-tagging, however, is not sufficient for the retrieval of relative clauses. At the same time, relative clauses are too frequent in a corpus of that size (i.e. approximately one million running words) to be extracted manually. Olofsson therefore considered automatic retrieval of the data. He (1981: 14) estimates that 95% of all occurrences of *which* are relative pronouns, but less than 20% of all uses of *that* are relatives. Even more problematic are zero relatives, as Olofsson (1981: 14) continues to point out:

> What deals the final and fatal blow, however, to the idea of letting the computer do the excerption of relative constructions, without pre-editing of the text material, is zero pronoun, which is out of reach as long as syntactic information is not included in the tape fed to the computer.

In the end, he resorted to the manual analysis of a representative sub-corpus, thus somewhat defeating the purpose of a computerized text corpus. Biber (1988), whose aim was to analyse a broad range of features in large corpora, only retrieved overtly marked relative clauses. Ball (1994: 297) discusses zero relatives as one of the problem cases for automatic analysis: "the identification of non-overt elements requires manual effort, a parsed corpus, or a robust parser, but there are considerable accuracy and coverage issues with currently available parsers [...]." She (1994: 301) even concludes that multi-factorial text analyses of the type that Biber (1988) conducted were 'premature' because they excluded zero relatives and thus

an important part of the possible envelope of variation. An early attempt at using POS-tagged data for present-day English was Lehmann (1997), a more recent one Huber (forthcoming). Lehmann uses 7 tag-based patterns to retrieve NP-NP combinations from a corpus of present-day English (tested against a retrieval strategy that relies on V sequences) and achieves 100% recall with this. Precision of his retrieval strategy is best for NP-NP strings where the second NP is a pronoun (Lehmann 1997: 185), but it is relatively low at 41%. It can be improved in a corpus that uses different tags for inflected verbs (ibid.: 187). A set of 9 constraints which limit the dataset further (pattern-matching algorithms, using regular expressions) brings recall down to 96% but precision up to 87% (ibid.: 191). However, Lehmann points out that the main problem with his study is that recall and precision were tested on the same data set that the retrieval strategy was developed on. Huber (forthcoming) only mentions that he retrieved sequences of two nominals that were manually post-edited but does not report findings on precision and recall for this retrieval procedure.

Recent developments in robust corpus annotation tools have made parsing of corpora much easier. This kind of syntactic annotation, in turn, makes retrieval of zero relatives a more realistic goal. Parser output has only been tested for Present-Day English data, so far. The objective of our paper is to test precision and recall of parser output for relative clauses in parsed historical data. 'Precision' is the technical term used for the targeting potential of a data retrieval procedure. It measures the percentage of true positives in the reported hits, i.e. how many of the automatically reported hits are correct. In other words, it concerns the proportion of 'false positives'. 'Recall' is the term used for the number of relevant strings that the search retrieves, i.e. how many (i.e. per cent) of the relative clauses in the text are found by automatic parsing. Precision usually decreases with an increase in recall (at the cost of manual post-editing), and vice versa. The parser was slightly adapted after a first analysis,

but the focus of this paper is not on parser improvements. We aim foremost to evaluate the usefulness and limitations of automatically retrieving relative clauses from a syntactically annotated corpus.

Our material has been taken from the scientific part of the ARCHER corpus (the forthcoming version 3.2), which we will briefly introduce in section 2 of this paper. We will detail the kinds of relatives we retrieved from the POS-tagged and parsed corpus in section 3. The evaluation of the parser output will present quantitative results on precision and recall as well as a discussion of examples that the parser failed to analyse correctly. We will conclude by making tentative suggestions for future approaches to automatic retrieval of relatives from annotated corpora.[1]

2. ARCHER and recent developments

Collaboration and extension of the original ARCHER corpus has been going on for several years.[2] In our paper we use the science texts of the corpus, including some American English scientific texts for all periods from 1700 onwards that were recently added to the corpus. Table 1 gives an overview of the data.

Table 1: Science texts in ARCHER 3.2 (number of words per sub-period)

	1700-49	1750-99	1800-49	1850-99	1900-49	1950-99
AmE	0	20,664	20,815	21,326	20,963	25,610
BrE	20,780	20,565	20,994	21,715	21,337	21,308

Our searches were based on a preliminary version of the forthcoming ARCHER 3.2 corpus which includes two additional files for the second half of the twentieth century in the American subpart of the corpus, hence the slightly larger subcorpus in this period.

The science part of the ARCHER corpus was annotated with a probabilistic parser (Pro3Gres) developed by Schneider (2008). The parser uses a dependency-based grammatical model close to Tesnière's Dependency Grammar conception (1959), combining a hand-written *competence* grammar and probabilistic *performance* disambiguation learnt from the Penn Treebank (Marcus, Santorini and Marcinkiewicz 1993). It was designed to cover the most frequent phenomena of standard Present-Day English grammar. It is fast (the BNC parses in under a day) and has been evaluated on several genres and varieties (Haverinen, Ginter, Pyysalo and Salakoski 2008, Lehmann and Schneider 2009). It is suitable for parsing different Englishes, as it is robust, so that its output is quite reliable on a number of English varieties (Schneider and Hundt 2009). An evaluation of the performance on subject, object and PP-attachment relations, using the GREVAL gold standard (Carroll, Minnen and Briscoe 2003) and 100 random sentences from the BNC, is given in Table 2.

Table 2: Performance of the Pro3Gres parser.

Performance on GREVAL (500 sentences)	Subject	Object	Noun-PP	Verb-PP
Precision	92%	89%	74%	72%
Recall	81%	84%	66%	84%
Performance on BNC (100 sentences)				
Precision	86%	87%	75%	89%
Recall	83%	88%	77%	70%

Precision and recall of relative clause parsing were also evaluated for PDE: Schneider (2008: 188) reports 91% precision and 68% recall for the anaphora of relative clause subjects (but this only concerns a subset of all possible relative clauses we are interested in here).

3. Defining the scope of relative clause retrieval

Relativizers in standard PDE include *who, whom, whose, which, that*, zero, plus relative adverbs *where, when, why, whence,*[3] *where* + Prep (*whereby*, etc.), plus other possibilities (*what, as*, etc.) some chiefly dialectal. Most of the first six typically relativize an NP and themselves constitute an NP,[4] though *whose* (always) and *which* (sometimes, though now only rarely) can function as determiners in the relative phrase:[5]

(1) In the $_1$paper $_1$*whose title is given above* the author has shown … (ARCHER, 1874mall.s6b)

(2) there arises … $_1$the necessity for a new supply of water … , $_1$*which necessity is met by* ... (ARCHER, 1886greg.s6a)

However, *whose* and *whom* are not part of the parser grammar. At the same time, a simple lexical search for these relativizers shows that they are very infrequent in our data, and we therefore feel confident in excluding them from the analysis.[6]

Initially, we had planned to retrieve reduced relative clauses as well. These would have been of interest from the methodological angle of this paper because, like zero relatives, they lack an overt relativizer and are thus a challenge for automatic data retrieval. However, they proved to be somewhat problematic from a theoretical point of view, as we will show. Consider these relative clauses, where the relativizer is subject of its clause and is followed by a form of *be* (examples 3a and 4a); there is a shorter form, more or less synonymous, in which both relativizer and *be* are missing (examples 3b and 4b):

(3) a. … those that survived $_1$the shock $_1$*which was occasioned by this sudden transition* … (ARCHER, 1874gunt.s6b)

b. ... those that survived the shock occasioned by this sudden transition ...

(4) a. Most of $_1$the studies $_1$*which are attempting to move to higher levels of complexity* ... (ARCHER, 1975macm.s8b)

b. Most of the studies attempting to move to higher levels of complexity ...

Some scholars accordingly describe the latter types, in which a participial phrase postmodifies a noun, as "reduced relatives" (e.g. Quirk, Greenbaum, Leech and Svartvik 1985: 418). Biber *et al.* explicitly compare them to "full relative clauses" (Biber, Johansson, Leech, Conrad and Finegan 1999: 631-632), which implies that they are reduced relative clauses, though Biber *et al*. retain the more cautious label "postmodifying participial clauses". Other scholars, e.g. Sag (1997: 433, 471-3), not only use the "reduced relative" label but go further and extend it to postmodifying patterns headed by something other than a participle (examples 5b and 6b):[7]

(5) a. She used an $_1$apparatus $_1$*which was similar in principle to* … (ARCHER, 1925dymo.s7b)

b. She used an apparatus similar in principle to…

(6) a. … by means of $_1$the visual image, $_1$*which was greatly out of focus* on account of the … (ARCHER, 1895keel.s6a)

b. … ?by means of the visual image greatly out of focus…

Quirk *et al*. say of certain types of postmodifying adjective that they "can be seen as reductions of relative clauses" (1985: 1294); examples would be

(7) This fact may probably lead to something <u>useful</u> hereafter. (ARCHER, 1791rush.s4a)

(8) This stomach possesses a property similar to that of the bladder ... (ARCHER, 1851dadd.s6a)

But while some post-nominal PPs can plausibly be regarded as reduced relative clauses (e.g. *out of focus* in example 6b), others cannot; *15 minutes* in (9) cannot be the antecedent of a relative clause:

(9) WE began to look for the first contact of Venus with the Sun, at least 15 minutes (*which were) before the time given by calculation (ARCHER, 1769west.s4a)

Discriminating between those that can and those that cannot is hard, both for humans and parsing programs. It seems, then, that to include putative reduced relative clauses would be to introduce a very fuzzy boundary to our dataset, as well as bringing in numerous examples lacking a relativizer. Huddleston and Pullum refuse to class them as relative clauses "since there is no possibility of them containing a relative phrase" (2002: 1265). But there are further reasons for caution. Quirk et al. observe a couple of properties of postmodifying participial clauses which distinguish them from relative clauses, including participial *-ing* forms that could not correspond to a progressive verb in a full relative clause (Quirk, Greenbaum, Leech and Svartvik 1985: 1263):

(10) a thick heavy syrup (*which is) resembling Melasses [*sic*] (ARCHER, 1791rush.s4a)

With non-restrictive postmodification there is the possibility of movement to initial position for the alleged reduced relatives (11b. and 11c.) but not for the full relative clauses (11 d.):

(11) a. … the visual $_1$image, $_1$*which was greatly out of focus*, … (ARCHER, 1895keel.s6a)

b. the visual image, greatly out of focus, …

c. greatly out of focus, the visual image …

d. *which was greatly out of focus, the visual image …

Furthermore, that very mobility "implies that nonfinite non-restrictive clauses are equivocal between adnominal and adverbial role" (Quirk, Greenbaum, Leech and Svartvik 1985: 1271). All in all, then, so-called reduced relative clauses are a rather dubious category. It remains true that postmodifying participial clauses and relative clauses can often be seen as linguistic alternatives. Because of the theoretical and practical problems outlined above, however, we decided to narrow our definition of the variable somewhat and exclude reduced relative clauses from our study.

Our study is part of a larger project on developments in noun phrase complexity in the late Modern period (Hundt, Denison and Schneider in prep.). We therefore confine attention to adnominal relative clauses, i.e. those with an NP as antecedent.[8] By focusing on adnominal relative clauses, we are looking at data sets that are comparable with previous studies and avoid unwanted statistical noise (see Sigley 1997: 37-40). The adnominal relative clauses we include in our study are introduced either by a standard[9] English *wh*-pronoun (i.e. *who* and *which*) or *that*; in addition, we also retrieved zero relatives automatically.

4. Evaluation of parsed data

We initially analysed for precision some concordances (AmE data from the 1700s) that were retrieved from the parsed corpora. After the parser had been adapted, we tested precision on a larger set of concordances across both BrE and AmE as well as across time. Note that we only tested for precision of relative clause parsing, not for precision of completely parsed sentences. The VP in the following example, for instance, was analysed wrongly in our output files with *be* as the head of the VP rather than *perforate*, but the relative clause was correctly identified:

(12) The effects of a yearly discharge of sap from the tree in improving and increasing the sap is demonstrated from the superior excellence of those $_1$trees *which have been perforated* in an hundred places, by a small wood-pecker which feeds upon the sap (ARCHER, 1791rush.s4a).

Ball (1994: 295) claims that recall is more difficult to assess in large corpora than precision: "it is generally impossible for the analyst to know what has been missed without analysing the entire corpus by hand." Such an extreme view tacitly assumes that it is impossible to extrapolate recall tested on a small set of texts to the corpus as a whole. While an exact measure of recall is indeed impossible to attain, and it is impossible anyway for other reasons like inter-annotator agreement, we would like to claim that testing recall in a subset of the corpus analysed automatically can give a good indication of recall for the corpus as a whole. The recall evaluation on the subset is as reliable as the subset is representative of the whole corpus. In addition to testing for recall of all adnominal relative clauses by analysing a subset of data we have also tested recall for the most frequent relativizer in scientific writing, namely *which*. As a first step we thus manually annotated all relatives (*who, which, that* and zero) in a subset of files and verified whether these had been identified by the parser by cross-checking the manually retrieved relative clauses with those listed in the concordances. The subset consisted of a total of 13 files in all, five texts from the 1700s (approx. 10,000 words) and 4 texts each for the 1800s and 1900s (i.e. approx. 16,000 words for the two subperiods together). As with precision, we only tested for recall of relative clauses, not for correctly parsed sentences. The antecedent in the following example was given as *what* (which could, arguably, be correct), but a much more obvious antecedent for the relative clause would be *ice*:[10]

(13) The ice-makers attended the pits usually before the Sun was above the horizon, and collected in baskets **what** was frozen, by pouring the whole contents of the pans into them and thereby retaining the $_{1}$ice, $_{1}$*which was daily conveyed to the grand receptacle or place of preservation* […] (ARCHER, 1775bark.s4b).

In another instance, the parser had wrongly picked out a relativizer (*who*) as the antecedent of another relativizer (*which*):

(14) A particularly favourable one has been afforded me lately through the kindness of $_{1}$Mr. GORDON, $_{1}$***who*** *has furnished me with considerable quantities of a* $_{2}$*fluid* obtained during the compression of oil gas, $_{2}$*of which I had some years since possessed small portions*, sufficient to excite great interest, but not to satisfy it. (ARCHER, 1825weav.s5b)

This was counted as a relevant example despite the fact that antecedent had not been correctly identified.[11] There are also examples where the verb of the relative clause was not identified correctly, and it would thus (in theory) be impossible to decide which relative clause was correctly parsed, as in the following example (where the parser misanalysed the inflectional ending *'d* as a reduced form of the verb *have* because it is part of the lexicon implemented in the parser):[12]

(15) An irregular gust of Wind blowing upon and shaking the Columns, was (I suppose) the Cause of that $_{1}$trembling, $_{1}$*which appear'd in the triangular Streams*, and the $_{2}$Cause also $_{2}$*which destroy'd that fine appearance of the Canopy*. (1720cote.s3b)

In this particular instance, two relative clauses with relativizer (*which*) and verb (*'d*) were retrieved from the corpus. In other words, the example did not prove a problem for the evaluation of recall.

In the following, we will first present the overall results on precision and recall and then turn to the discussion of individual examples.

4.1 Quantitative results

Overall, the initial results for precision were 83.5% for relatives introduced by *who/which/that* and 18.5% for zero relatives; recall was 43%. The parser was subsequently adapted with the help of the comments in the concordances.

The parser grammar was adapted in several ways, some directly addressing shortcomings in the analysis of relative clauses, some generally improving performance, and some addressing phenomena that are more frequent in previous stages of English. Perhaps surprisingly, none of the changes addresses features that no longer exist in PDE. Each change turned out to be a general parser improvement. The parser was adapted in the following five ways:

First, *whom* and *whose* were added as relativizers, correcting a shortcoming of the parser that is relevant for earlier stages of English in general (even though they occurred infrequently in our scientific data).

Second, the parser grammar was adjusted so that pied-piping constructions are now parsed correctly. Previously, relative pronouns inside a prepositional phrase were explicitly disallowed to post-modify NPs, as is the case in pied-piping constructions. This poses problems for sentences such as the following example:

(16) He informed me that in his journey from Passy to Havre de Grace, last summer, he found the $_1$country $_1$*through which he travelled*, unusually sickly with fevers. (ARCHER, 1786rush.s4a)

In the parser output before adjustment, the relative pronoun *which* directly attached to *country* instead of the preposition *through*. Pied-piping constructions are more frequent in historical data. This adaptation generally improves the parser output. For ARCHER, it was the single most beneficial adaptation made.

Third, candidates for relative pronouns acting as objects are only allowed to attach if no overt object exists. In the following sentence, the original parse had the verb *represent* taking two objects, namely *mountains* and the relative pronoun object *that*.

(17) He gives the preference to the Gregorian, and mentions as a principal defect of the Cassegrain telescope, that it represents the mountains in the moon as vallies, and the contrary. (ARCHER, 1786ritt.s4a)

Valency checks in principle forbid several objects, but long-range objects such as relative pronouns were not sufficiently subjected to these checks. After the correction, this example is no longer incorrectly reported as a relative clause, but correctly as a subordinate clause. This adaptation, in addition to improving performance on historical English, also minimally improves parser performance on PDE.

Fourth, a list has been created of words that are licensed to be complementizers. In the Penn Treebank tagset, which the parser uses, the tag IN is used for complementizers, prepositions and relativizers. In rare cases, relative pronouns like *which* were analysed as complementizers. However, the ambiguity between complementizer and relative pronoun, particularly of *that*, is still a major source of errors, which brings us to our next point.

Fifth, the class of nouns that can introduce subordinate clauses is now learnt from the Penn Treebank, while previously a small, closed list was used. This adaptation only leads to minimal improvements, as the ambiguity depends on semantics and verb valency in ways that the parser does not sufficiently respect. In *The suggestion that we should follow* we probably

298 have a relative clause, while in *The suggestion that we should go* we probably have a
299 subordinated clause. In real world sentences, which are usually more complex than invented
300 examples, attachment ambiguities and the complementizer/relative pronoun ambiguity
301 combine forces. In the following example, the parser attached *that* to *part* as relativizer
302 instead of attaching it to *consequence* as a complementizer.

303 (18) It is in *consequence* of the sap of these trees being equally diffused through every
304 *part* of them, *that* they live three years after they are girdled [...]. (ARCHER,
305 1791rush.s4a)

306 In the next example, the parser attached *that* to *tube* as relativizer instead of to *found* as
307 complementizer:

308 (19) And accordingly I *found*, upon taking out both of the glasses, and looking through
309 the open *tube*, *that* the hearth appeared as perfectly, and as constantly in its
310 unnatural state by reflected light [...]. (ARCHER, 1786ritt.s4a)

311 Parser adaptation led to improved precision (85% for *wh-/that* relatives and 28% for zero
312 relatives) and recall (53%) for the 1700s American English data. Table 3 below gives a more
313 detailed summary of the results for the two varieties and across time.

314 Table 3: Precision (after parser adjustment)[13]

AmE	1700s	1800s	1900s
wh-/that	85%	83%	88%
zero	28%	11%	11%
BrE			
wh-/that	86%	82%	82%
zero	20%	24%	0%

pooled results (AmE and BrE)			
wh-/that	86%	83%	86%
zero	23%	18%	5%

As expected, the parser performs best on overtly marked relative clauses. Note that the parser was adjusted on the basis of the 1700s data and improvement led to better performance in this period, also for zero relatives. Precision for overtly marked relative clauses is also quite high for the 1800s and 1900s, but for zero relatives the figures remain quite low. We doubt, however, whether parser adjustment to data retrieved for the latter two sub-periods would substantially improve the precision of zero-relative retrieval.

Recall is much lower, overall, than precision for overtly marked relative clauses, and it is better for the twentieth-century data than for the earlier periods, as Table 4 shows:

Table 4: Recall (after parser adjustment; both varieties)

sub-period	correctly identified relatives	recall
1700s	40 out of 92	43%
1800s	29 out of 71	41%
1900s	40 out of 76	53%

Breaking down recall errors by relative pronoun reveals that, quantitatively, the impact of missed zero or *that* relatives is small. In fact, the majority of relatives that were not retrieved automatically are *wh*-relatives (see Table 5), the most commonly used relativizer in the science part of ARCHER overall:

Table 5: Missed relative clauses by relativizer

sub-period	*that*	*which/who*	zero	total
1700s	4	46	2	52
1800s	2	39	1	42
1900s	9	25	0	36

In absolute terms, then, the relative clauses that are particularly difficult to retrieve automatically (i.e. *that*- and *zero*-relatives) turn out to perform quite well in science texts with respect to recall. In absolute terms, the easy-to-retrieve *wh*-relativizers are missed more often than the ones that are difficult to retrieve automatically. As recall of relativizer *which* seemed especially low (probably also because *which* is a particularly frequent relativizer in our data), we decided to retrieve all instances of *which* from the bare-text version of our corpus, manually delete all instances that were not adnominal relative clauses, in order to test more widely for recall of this important relativization strategy in historical scientific writing. The results are given in Table 6.

Table 6: Recall for adnominal relatives introduced by *which*
(a = automatically retrieved, m = manually retrieved)[14]

	1700s			1800s			1900s		
	a	m	recall	a	m	recall	a	m	recall
AmE	110	181	61%	160	367	44%	127	224	57%
BrE (50s-99s)	103	174	59%	87	180	48%	59	111	53%

Even though relative clauses introduced by *which* make up a large part of the missed relative clauses numerically, recall for these relative clauses in a larger section of the corpus is above that for all types of relative clauses considered in our paper.

A closer look at individual examples from the concordances for both precision and recall will show where some of the potential problems for parsing lie.

4.2 Discussion: Precision

One of the reasons why precision with zero relatives is so low in the historical data is that fronting of objects is still common in earlier texts. Examples (20)-(24) are typical instances of false positives (false antecedents are given in italics, erroneously retrieved relative clauses are underlined):

(20) *Similar relics* I have found in the stomach of the pneumora and gryllus virridissimus. (ARCHER, 1825kidd.s5b)

(21) *This conclusion* I deduced from the fact, discovered by DELAROCHE, that invisible caloric freely permeates very thin plates of glass, in the same manner as light, but that it is completely intercepted by thicker plates. (ARCHER, 1825pond.s5b)

(22) But *this* I am unable to do; as I will show, by stating a circumstance rather deserving of attention. (ARCHER, 1825wood.s5b)

(23) *This irregularity of curve* I consider to be the most vexatious fault a mirror can have. (ARCHER, 1874lass.s6b)

(24) *The Places of these two Stars* I have not yet observed. (ARCHER, 1724brad.s3b)

The parser identifies these fronted objects as antecedents of a zero relative. This is a further task for future parser adaptation. However, the vast majority of false positives defy neat classification. The following is just one example of a real problem case:

(25) The present research forms part of a wider investigation of terrestrial magnetism, the main *object* of which is the study of certain electrical

phenomena that are associated with solar emissions absorbed in the upper atmosphere, and with the systematic motions of the upper atmosphere. (ARCHER, 1925chal.s7b)

The parser identified *object* as the antecedent of the relative clause, *associate* as its verb and *which* as the relativizer, but of course the verb is *is* and the antecedent *investigation*, in a structure which the Huddleston, Pullum & Peterson (2002: 1040-1042) would describe as involving (twofold) upward percolation.

4.3 Recall

As far as recall is concerned, sentence length and complexity in early scientific writing pose obvious problems. Ambiguity increases exponentially when sentences are long. Example (26) illustrates the problem of sentence length; the sentence contains three adnominal relative clauses, two of which were retrieved automatically. (We enclose undetected relative clauses in square brackets):

(26) The $_1$opinion, therefore, $_1$*which I have formed from what I have hitherto seen is*, that the boiled and common water differ from one another in this respect; that whereas the common water, when exposed in a state of tranquillity to $_2$air $_2$*that is a few degrees colder than the freezing point* may easily be cooled to the degree of such air, and still continue perfectly fluid, provided it still remain undisturbed; the boiled water, on the contrary, can not be preserved fluid in these circumstances; but when cooled down to the freezing point, if we attempt to make it in the least colder, a part of it is immediately changed into ice; after which, by the continued action of the cold air upon it, more ice is formed in it every moment, until the whole of it be gradually congealed before it can become

as cold as the $_{3}$air [*that surrounds it*.] (ARCHER, 1775blac.s4b)

Sentence complexity is generally a problem because it inflates ambiguity. In principle, the parser can cope with any depth of nesting or stacking, but because this increases syntactic ambiguity, nesting and stacking also pose problems for parsing. In a configuration where two relative clauses follow a noun phrase, for instance, the second relative clause can potentially modify any noun phrase in the first relative clause or even the initial NP. Thus, in the following example the first relative clause (introduced by *that*) was identified correctly by the parser, but the second one – separated by a series of other subordinate clauses – was not:

(27) The first $_{1}$rudiments of this art $_{1a}$*that I acquired* was from the two Hunters, known through all Europe for their superior skill in anatomy, and acting as practical dissector to the celebrated doctors Colignon and Smith, professors of anatomy in the universities of Cambridge and Oxford, $_{1b}$[*which I further improved by practice at Paris with Mons, Süe*]$_{2}$ to whom I am wholly indebted for my knowledge of anatomical preparations in wax. (ARCHER, 17??morg.s4a)

The bracketed *that*-relative in the following example is nested within another relative clause and might have been missed by the parser for that reason:

(28) several other $_{1}$Striae were discharged from behind the dark Basis, $_{1}$*which intersecting with* $_{2}$*others*, $_{2}$[*that at the same time arose about the East and West Points,*] *form'd in the Zenith*, or rather 6 or 8 degrees to the South thereof, a second much more elegant and surprizing than the former, and indeed than $_{3}$any thing $_{3}$*that had yet appeared*: (ARCHER, 1721lang.s3b)

The following is a particularly complex example with four relative clauses, of which one (in brackets) was not identified by the parser:

(29) The process results in the production of a $_{1}$form $_{1a}$*which I proposed to call the*

Planula, but $_{1b}$[*which Professor HAECKEL has better termed the Gastrula*], reserving the former name for a $_{2}$condition of the Gastrula $_{2a}$*which sometimes presents itself* $_{2b}$*in which there is no aperture of invagination*. (ARCHER, 1874lank.s6b)

Particularly difficult to evaluate were cases in which an antecedent was followed by two relative clauses that were introduced by the same relativizer each time. The probabilistic model of the parser includes distance (measured in chunks) which often (but not always) means that close attachments are preferred over more distant ones. The following example occurred in our list of automatically retrieved relative clauses only once with the verb *be* as the verb of the relative clause; in other words, the second relative clause, introduced by *through which*, was not parsed correctly (and thus not retrieved):

(30) That all those $_{1}$parts of any animal Body, $_{1a}$*which are vascular*, or $_{1b}$[*through which any Fluid passeth*,] from the intestines to the minutest Fibre, are the seat of medicine's Operation. (ARCHER, 1720cote.s3b)

At other times, distance from the antecedent did not cause a problem for the parser. The following example was parsed correctly, for instance:

(31) As to the Knife, it was not the Blade, but the $_{1}$Haft, and the $_{2}$Hinge $_{2}$*that goes into it*, $_{1}$*which was found shiver'd in Pieces*. (ARCHER, 1725nett.s3b)

Note that in example (29) above, the relative clause headed by a preposition (subscript 2b) was parsed correctly. The missed relative clauses in both (29) and (30) are preceded by a co-ordinating conjunction. Co-ordination creates serious problems for the parser, since the conjunction can combine elements at any level (NP, VP, clause). Statistical models are not of much help here, as there is typically no lexico-grammatical preference. A third example where the second relative clause was missed seems to confirm this:

(32) A young Scotch $_{1}$fir, $_{1a}$*which had two compleat shoots and a third growing*, and$_{_1b}$ [*which consequently was in its third year*], was put into the cold $_{2}$mixture $_{2}$*which was between 15 and 17*. (ARCHER, 1775hunt.s4b)

But missing coordinating conjunctions between double relative clauses, likewise, pose a problem for automatic retrieval of relative clauses; in the following example, only the first but not the second relative clause (in brackets) was retrieved in our parser-based approach:

(33) The kind of $_{1}$preparations of those $_{2}$parts of the animal body $_{2}$[*which admit of it*] $_{1}$*that I now propose to explain*, namely by injection and corrosion, exceeds in beauty, nicety and usefulness, that which is commonly called dissection. (ARCHER, 1786morg.s4a.txt)

Rissanen (1984: 424, exx 4 and 5) points out that relative clauses which are close to their antecedent are more likely to be introduced by zero or *that* and that, with growing distance between antecedent and relativizer, the likelihood increases that the semantically more transparent *wh*-relativizers will be used.[15] This is illustrated by the following example of a double relative clause:

(34) Two left off feeding; these I placed on the $_{1}$racks $_{1a}$*ø I had made*, $_{1b}$*which I fixed in glass bottles* to prevent the worms from getting off: [...] (ARCHER, 1769bart.s4a)

But we also came across an interesting counter-example to the general principle that closeness to the antecedent favours zero and *that*-relatives and that with growing distance, the more explicit *wh*-relatives are preferred. In example (33) above, the hierarchically second relative clause is introduced by a *wh*-relativizer but the hierarchically first relative clause (i.e. the one with the antecedent *the kind of preparations*) actually follows it and is introduced by the less explicit *that*. In other words, a more semantically transparent relative construction is

nested within a long-distance head-relativizer sequence which, surprisingly, has *that* as relative pronoun.

Punctuation in relative clauses in previous stages of the language was somewhat different from the conventions in PDE (see Denison and Hundt in prep.). The prescriptive rule of separating only non-restrictive relative clauses by a comma was not firmly in place yet. Even zero relative clauses, which are always restrictive, are sometimes set off from the remaining text by commas, and therefore the parser grammar has missed the following relevant instance (it only has a rule for zero relatives without a comma):

(35) The first $_1$Experiment, $_1$[*ø I have to offer to your Observation at present*,] is made on the New England Cedar, or rather Juniper, grafted on the Virginia; and what is remarkable in it, is, That the $_2$Branch, $_2$*which is grafted*, is left several Inches below the Grafting, which part continues growing as well as the upper Part above the Grafting. (ARCHER, 1724fair.s3b)

In addition to changes in punctuation, changes in relative clause structure have also caused problems for the parser. Retrieving relatives that are obsolete (or obsolescent) in PDE would be a case in point. In the following example, the antecedent *calculations* is repeated after the relativizer:

(36) In the report of March, 1812, page 9, the commissioners gave $_1$calculations on the expense of conveyance of canals, $_1$*which calculations were drawn from the experience acquired on canals in England*, as to the quantity of work that two horses and three men could do in eight hours; (ARCHER, 1814morr.s5a)

A similar 20th-century example does not have a repeated antecedent, and indeed it is not clear whether the relative is adnominal at all:

(37) In 1849, Carpenter began the study of the wall structure of Foraminifera, restricted largely to the large, calcareous forms, *which work was completed in his important "Introduction to the Study of Foraminifera", in 1862*. (ARCHER, 1928gall.s7a)

Adding a repetition rule to the parser grammar for earlier historical texts might even have a negative effect on parser performance.

The following example is no longer possible in PDE because the antecedent would have to be nominal and the relativizer would have to be preceded by a preposition (i.e. *the promotion of which*):[16]

(38) The manufacturing interest, *to promote which* is one of the objects of the society, is a subject of much importance, and would furnish matter far beyond the limits of an address. (ARCHER, 1823adam.s5a)

Finally, punctuation might again account for the fact that the following restrictive relative headed by *which* was not retrieved by the parser:

(39) I was at Gibraltar when this happen'd, where I saw above 100 of the Butts of that $_1$Cargo of Brandy, $_1$[*which were sent thither from Tangier*]; I likewise spoke with the $_2$Captain of the Dutch Ship, $_2$*who told the Governor, myself, and many others where his Vessel sunk*; and her rising afterwards at Tangier, appear'd very unaccountable to us, as it does to me to this Day; for there's no Doubt but the Ship sunk where the Dutchman told us, since the $_3$Spaniards from the Land, $_3$*who saw it*, confirm'd it to us. (ARCHER, 1724fair.s3b)

In this example, the correct attachment could have been found by the grammar, but apparently it was judged to be less likely.

Sentential relatives are also a problem for automatic retrieval: the distinction between sentential and adnominal relative clauses is difficult because it is more semantic than

syntactic, and the parser therefore always attaches relative pronouns to a noun, which means that all sentential relatives get an incorrect parser analysis and have to be removed manually from the parser output. In the following example, *which* is attached to *glass* instead of to the clause introduced by *from*. As it is a relatively long sentence, which is typical for the period, attachment ambiguities multiply.

(40) When I expected the flies were near coming out, I tacked coarse cloths up against the windows on the inside, not only to darken the room, but also for the flies to settle on, and to prevent them, in attempting to make their escape, from $_1$beating their legs and wings to pieces against the glass, $_1$*which I found to be the case last year*, and which it is probable, prevented their copulating. (ARCHER, 1769bart.s4a)

Another problem for the parser is the expression *so ... that*: it is not included in the parser, partly because few multi-word expressions are recognised, and partly because PDE newspaper and science texts, on which the parser was developed, do not contain many *so ... that* constructions. In ARCHER, however, they frequently cause parser errors. In the following example, *that* gets attached to *matter*.

(41) My answer gave him so much satisfaction in the matter, that he immediately sent his orders to his correspondent in London, to procure the instruments. (ARCHER, 1769west.s4a)

The parser treated semi-colons as a sentence boundary, yet relative clauses in early texts are occasionally punctuated off by a semi-colon, as in the following example:

(42) between N.W. by North, and W.N.West, we found the Representation of a very bright Crepusculum, such as $_1$that *which appears about 20 Minutes after Sun-set*;

[1]from *which arose several very large Beams of Light*, not exactly erect towards the Vertex, but somewhat declining to the South; […] (ARCHER, 1721lang.s3b)

This detail had not been changed in the first adaptation of the parser. Future parser adaptation in this direction (also allowing for relative clauses to be separated from the previous clause by a colon or to occur in parentheses) is expected to improve recall of non-restrictive *wh*-relative clauses, i.e. those that tended to be set off by other punctuation than a comma in our earlier texts on a preliminary search.

5. Conclusion

The aim of our paper was to evaluate the possibilities and limitations of retrieving relative clauses from a parsed corpus of historical English. We found that, initially, precision was 83.5% for overtly marked relatives and but as low as 18.5% for zero relatives; recall was 43%. Parser adjustment improved precision for overtly marked relatives somewhat. For zero relatives, the precision could be improved to 28% for American scientific texts of the 1700s, but was found to be as low as 0% for the British texts of the 1900s. We found that the parser identified fronted objects as antecedents of zero relatives. Future parser adaptation is therefore likely to slightly improve the accuracy of the parser output in this area and thus improve precision. The main problem will be that the majority of false positives of zero relative clauses defy easy classification and thus do not substantially contribute to parser improvement. Despite adaptations, many seemingly simple problems with parsing historical data persist, such as the complementizer ambiguity (see the discussion of examples 18 and 19 above), which poses a problem for precision. With respect to recall, the qualitative analysis of the parser output showed that sentence length and complexity (including nesting) were the most obvious problems for parser failure.

One of the potential advantages of working with parsed data is that it allows one to retrieve zero relatives. These relative clauses remain problematic because precision for zero relatives is so low. Recall for zero relatives in our scientific data, on the other hand, was quite good, but this was mainly because zero relatives are infrequent in this kind of data. They are thus more likely to be a problem for more informal written text types and spoken data and future work on parser adaptation should therefore include these text types. A combination of the tag-based approach described in Lehmann (1997) with the parser-based approach taken in this paper might improve both precision and recall of automatically retrieved zero relatives.

565

[1] More details on adapting the parser to historical texts are given in Schneider (2011).

[2] For the development of the ARCHER corpus, see Yáñez-Bouza (2011).

[3] Huddleston, Pullum and Peterson (2002: 1051) include *while* in this list.

[4] Note that the NP belongs to a PP in the case of pied piping (an instance of 'upward percolation' in the terminology of Huddleston, Pullum and Peterson (2002: 1040)).

[5] Relative clauses and their antecedents are marked by subscripts throughout this paper.

[6] On the choice of *who* vs. *whom*, see for instance Tieken-Boon van Ostade (1990) or Schneider (1992a, 1992b).

[7] In fact in early transformational grammar there was a move to derive all attributive adjectives, even premodifying ones, from predicative adjectives in relative clauses (e.g. … an apparatus which is/was similar ⇨ an apparatus similar ⇨ a similar apparatus) (see e.g. Culicover 1982).

[8] Sentential relative clauses (e.g. *He manages to swim the whole length of the pool, which amazes me*) and free relatives (e.g. *We will be working on this on Saturday, which will be nice* or *He always gets what he wants*) are examples of relative clauses that are not adnominal. They are likely to have undergone significant changes in the late Modern English period, too, but this will have to be the topic of future research.

[9] It is highly unlikely that non-standard relativizers like *what* or *as* would be used in scientific English.

[10] Incorrectly identified antecedents are highlighted in bold throughout the paper.

[11] Note that the parser failed to identify the relative clause with *Mr. Gordon* as antecedent!

[12] As part of further improvement in the annotation of ARCHER, spelling variation in the corpus was normalized with VARD (see Schneider 2011). Running the parser over PDE spelling variants significantly improves the overall performance of the parser.

[13] For a table with frequencies of positives and false positives, see the appendix.

[14] Note that the figures for the BrE part of the corpus are based on the data from the second half of the century only.

[15] See also Quirk (1957: 106 n.8) on educated spoken English: "The preponderance of *wh-* is at its greatest in the cases of double restriction (as in 'there are certain activities which are not scientific which are very important to the human race'), where the second clause has *wh-* 15 times as against *that* 5 times and no examples of zero."

[16] An alternative pattern using *promote* as a verb would be *The manufacturing interest, which it is one of the objects of society to promote*…

References

Ball, C.N. (1994) 'Automated text analysis: Cautionary tales'. *Literary and Linguistic Computing* **9**(4): 295-302.

Ball, C.N. (1996) 'A diachronic study of relative markers in spoken and written English'. *Language Variation and Change* **8**(227-258.

Biber, D. (1988) *Variation across speech and writing*. (Cambridge, etc: Cambridge University Press).

Biber, D. and Clark, V. (2002) 'Historical shifts in modification patterns with complex noun phrase structures' in T. Fanego *et al.* (eds.), *Historical shifts in modification patterns with complex noun phrase structures*. Amsterdam and Philadelphia PA: John Benjamins, pp. 43-66.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999) *Longman grammar of spoken and written English*. London: Longman.

Carroll, J., Minnen, G. and Briscoe, E. (2003) 'Parser evaluation: Using a grammatical relation annotation scheme' in A. Abeillé (ed.), *Parser evaluation: Using a grammatical relation annotation scheme*. Dordrecht: Kluwer, pp. 299-316.

Culicover, P.W. (1982) *Syntax* (2nd Edition). (New York: Academic Press).

Denison, D. and Hundt, M. (in prep.) 'Defining relatives'.

Haverinen, K., Ginter, F., Pyysalo, S., Salakoski, T. (2008) 'Accurate conversion of dependency parses: targeting the Stanford scheme.' *Proceedings of third international symposium on semantic mining in biomedicine* (SMBM 2008), Turku, Finland, pp. 133-136.

Huddleston, R. (2002) 'Non-finite and verbless clauses' in R. Huddleston & G.K. Pullum (eds.), *Non-finite and verbless clauses*. Cambridge: Cambridge University Press, pp. 1171-1271.

Huddleston, R., Pullum, G.K. and Peterson, P. (2002) 'Relative constructions and unbounded dependencies' in R. Huddleston & G.K. Pullum (eds.), *Relative constructions and unbounded dependencies*. Cambridge: Cambridge University Press, pp. 1031-1096.

Hundt, M., Denison, D. and Schneider, G. (in prep.) 'Relative complexity in scientific discourse'. *English Language and Linguistics*

Johansson, C. (2006) 'Relativizers in nineteenth-century English' in M. Kyto *et al.* (eds.), *Relativizers in nineteenth-century English*. Cambridge: Cambridge University Press, pp. 136-182.

Lehmann, H.M. (1997) 'Retrieval of zero elements in a computerised corpus' in M. Ljung (ed.), *Retrieval of zero elements in a computerised corpus*. Amsterdam: Rodopi, pp. 179-194.

Lehmann, H.M. (2002) 'Zero subject relative constructions in American and British English' in P. Peters *et al.* (eds.), *Zero subject relative constructions in American and British English*. Amsterdam and New York: Rodopi, pp. 163-177.

Lehmann, H.M. and G. Schneider. (2009) 'Parser-based analysis of syntax-lexis interaction' in A.H. Jucker *et al.* (eds.), *Parser-based analysis of syntax-lexis interaction*. Amsterdam and New York: Rodopi, pp. 477-502.

Marcus, M.P., Santorini, B. and Marcinkiewicz, M.A. (1993) 'Building a large annotated corpus of English: The Penn Treebank'. *Computational Linguistics* **19**(2): 313-330.

612 **Olofsson, A.** (1981) *Relative junctions in written American English*. (Gothenburg: Acta
613 Universitatis Gothoburgensis).
614 **Quirk, R.** (1957) 'Relative clauses in educated spoken English'. *English Studies* **38**(97-109.
615 **Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J.** (1985) *A comprehensive grammar*
616 *of the English language*. (London and New York: Longman).
617 **Rissanen, M.** (1984) 'The choice of relative pronouns in 17th century American English' in J.
618 Fisiak (ed.), *The choice of relative pronouns in 17th century American English*. Paris
619 and The Hague: Mouton, pp. 417-435.
620 **Sag, I.A.** (1997) 'English relative clause constructions'. *Journal of Linguistics* **33**(431-483.
621 **Schneider, E.W.** (1992a) 'Who(m)? Case marking of wh-pronouns in written British and
622 American English' in G. Leitner (ed.), *Who(m)? Case marking of wh-pronouns in*
623 *written British and American English*. Berlin and New York: Mouton de Gruyter, pp.
624 231-245.
625 **Schneider, E.W.** (1992b) 'Who(m)? Constraints on the loss of case marking of wh-pronouns
626 in the English of Shakespeare and other poets of the Early Modern English period' in
627 M. Rissanen *et al.* (eds.), *Who(m)? Constraints on the loss of case marking of wh-*
628 *pronouns in the English of Shakespeare and other poets of the Early Modern English*
629 *period*. Berlin and New York: Mouton de Gruyter, pp. 437-452.
630 **Schneider, G.** (2008) 'Hybrid long-distance functional dependency parsing'. PhD, University
631 of Zürich.
632 **Schneider, G.** (2011) 'Adapting a parser to historical English'. Paper presented at the
633 Helsinki Corpus Festival, September 27 - October 2, 2011.
634 **Schneider, G. and M. Hundt.** (2009) 'Using a parser as a heuristic tool for the description of
635 New Englishes'. Paper presented at CL 2009, Liverpool.

Sigley, R. (1997) 'Choosing your relatives: Relative clauses in New Zealand English'. PhD, Victoria University.

Tesnière, L. (1959) *Eléments de syntaxe structurale*. (Paris: Librairie Klincksieck).

Tieken-Boon van Ostade, I. (1990) 'Betsy Sheridan: Fettered by grammatical rules?'. *Leuvense Bijdragen* **79**(79-90.

Yáñez Bouza, N. (2011) 'ARCHER past and present (1990-2010)'. *ICAME Journal* **35**(205-236.

646 Appendix

647 Table: Frequencies of relative clauses after parser-adjustment

AmE	1700s			1800s			1900s		
wh-/that	192			232			238		
positives – false positives – ambiguous	164	28	0	192	39	1	210	27	1
zero	72			72			46		
positives – false positives	20	52		8	64		5	41	
BrE									
wh-/that	301			305			176		
positives – false positives – ambiguous	260	42	0	252	53	0	144	31	1
zero	132			91			48		
positives – false positives	26	106		22	69		0	48	
both varieties									
wh-/that	493			537			414		
positives – false positives – ambiguous	424	70	0	444	92	1	354	58	2
zero	204			163			94		
positives – false positives	46	158		30	133		5	89	

648